CCOS TC Kickoff Meeting
Cluster Analysis for CCOS Domain

Ahmet Palazoglu (P.I.)
Scott Beaver
Swathi Pakalapati

University of California, Davis
Department of Chemical Engineering & Materials Science

June 20, 2006

Overview

- Introduction to Cluster Analysis
  - 2 Types of Cluster Analysis for Air Quality Applications
    - Daily maximum 8-hr [O_3]
    - Hourly ground-level wind field
- Cluster Analysis for CCOS Domain Project
  - Project Work Plan
  - Recent Progress
  - Information needed from CCOS and Districts

What is Cluster Analysis?

- *Unsupervised* statistical methods to determine recurring patterns from a set of observations.
  - Require no advanced knowledge of how patterns manifest themselves in data.
- Important
  - Representative input data
    - Spatial field of single parameter– O_3 & wind fields
    - Monitoring station network – must represent physical domain
    - Station selection – require aid from district staff
  - Appropriate statistical model
    - Different models identify different patterns/physical processes
    - Poor model choice might be misleading

2 Cluster Analyses for Air Quality

- General cluster analysis framework
  - Performed for set of days spanning a number of years
  - Consider simultaneous, spatially distributed measurements
  - Days are labeled forming clusters of similar days
- 2 models for specific air quality parameters
  - Daily maximum 8-hr $[O_3]$
    - Applied to a disjoint set of days experiencing high ozone levels (e.g. exceedance days for 1996-03)
    - Reveal recurring mechanisms resulting in exceedances
  - Hourly ground-level wind field (speed and direction)
    - Applied to continuous measurements for entire ozone season (e.g. each day from 1 June to 30 September, 1996-03)
    - Finds surrogate meteorological patterns and associated ozone response

Cluster Analysis for CCOS Domain

- Independent cluster analyses for 6 basins
  - Bay Area (BA)
  - Sacramento Valley (SV)
  - San Joaquin Valley (SJV)– North, Central, South
  - Mountain Counties (MC)
- Study Period
  - Ozone season only (1 June – 30 September)
  - Core period 1996—2004
  - 2 historical years before 1996
    - 1985 & 1990 ?
    - Should consider El Nino effect so results reflect emissions reductions?
  - Recent year 2005

Work Plan

- Phase I / Year I
  - MATLAB Toolbox for computations and visualization
    - Completed for Phase I
  - Wind field clustering for 6 basins
    - Completed for BA years 1996—2003
    - Require data for other basins/years
  - 8-hr $[O_3]$ clustering for 6 basins
    - Expanding on previous study for BA
    - Require data for other basins
- Phase II / Year II
  - Synoptic scale dynamics and inter-annual trends
  - Synopsis of CCOS domain air quality meteorology – to determine weather patterns affecting multiple basins

Recent Progress

- Database received for 1980–2004
  - Must select air quality and met. stations for each basin
- Develop MATLAB Toolbox
  - Computation & Visualization for Phase I complete
  - BA analysis as template for other air basins
- Revising BA Analysis
  - “Correlation” vs. “Euclidean” metrics for O_3 clustering
  - “Lowering of standard” for BA O_3 clustering

Correlation vs. Euclidean Metrics

- Original BA O_3 clustering
  - *k*-means clustering with Euclidean metric
    - severity of regional ozone levels
- Revised BA O_3 clustering
  - *k*-means clustering with Correlation metric
    - spatial variation of episode locations
  - Metric provides alternative perspective
    - Can identify previously undetected mechanisms?

Lowering of Standard

- Original BA O_3 clustering
  - 8-hr NAAQS exceedance days
    - Daily max. $[O_3] > 84$ ppb at any station
    - Only 63 episode days in 8 years
- Revised BA O_3 clustering
  - Include more days to gain broader perspective
    - Daily max. $[O_3] > 70$ ppb at any BA station
      - Different threshold for each air basin
    - Include all 1-hr and 8-hr exceedance days?
  - Have data for 18/22 BA stations from previous analysis
    - Multiple monitors in database for 4 unknown, urban sites
      - San Francisco, Oakland, San Jose, San Jose E.
      - Population exposure vs. maximum concentration?

Information Needed

- BAAQMD
  - AIRS codes for remaining 4/22 air quality stations
- Other Districts
  - Confirm air quality station network
  - Begin selection for met network
- CCOS
  - Coordinate efforts from districts
  - Obtain new data for 2005
  - Select pre-1996 benchmark years (1985 & 1990 ???)

Summary and Immediate Work

- Completed
  - MATLAB Toolbox for Phase I
  - Most of BA analysis
- In Progress
  - Refinement of O_3 clustering algorithm
    - Testing on BA data set
  - Obtaining data for other basins (SJV, SV, MC)
  - MATLAB Toolbox for Phase II
- Information needed
  - Historical years to include in study period
  - Confirm air quality networks for cluster analysis
  - Select meteorological monitoring network with representative wind data for desired study period

Links to Published Studies

- Daily maximum 8-hr [O_3] clustering
  - Beaver, S. and Palazoglu, A., 2006: A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area. *Atmos. Environ.*, 40, 713—725.
- Continuous, hourly ground-level wind field clustering
  - Beaver, S. and Palazoglu, A., 2006: Cluster analysis of hourly wind measurements to reveal synoptic regimes affecting air quality. In press *J. Applied Meteor*.